Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes

Dimauro C., Cellesi M., Steri R., Gaspa G., Sorbolini S. , Stella A. and Macciotta, N. P. P.

Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. *2013. Anim Genet, 44: 377-382. doi:10.1111/age.12021*

Running head: Bovine breed assignment and traceability

Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes

C. Dimauro,[*] M. Cellesi,[*] R. Steri,[*] G. Gaspa,[*] S. Sorbolini,[*] A. Stella,[†] N. P. P. Macciotta[*]

[*]Dipartimento di Agraria, Università di Sassari, Via De Nicola 9, 07100 Sassari, Italy

[†]Istituto di biologia e biotecnologia agraria CNR, Milano, 20133 Milano, Italy

Address for correspondence

C. Dimauro, Dipartimento di Agraria, University of Sassari, Via De Nicola 9, 07100 Sassari, Italy

Fax: +39079229302; telephone number +39070229298

E-mail: dimauro@uniss.it

Summary

Several market researches have shown that consumers are primarily concerned with the provenance of the food they eat. Among the available identification methods, only DNA-based techniques appear able to completely prevent from frauds. In this paper, a new method to discriminate among different bovine breeds and assign new individuals to groups was developed. Bulls of three cattle breeds farmed in Italy, Holstein, Brown and Simmental, were genotyped by using the 50K SNP Illumina BeadChip. The multivariate canonical discriminant analysis was used to discriminate among breeds whereas, the discriminant analysis was used to assign new observations The method was able to completely identify the three groups already at chromosome level. Moreover, a genome wide analysis developed by using 340 linearly independent SNPs yielded a significant separation among groups. Using the reduced set of markers, the discriminant analysis was able to assign 30 independent individuals to the proper breed. Finally, a set of 48 high discriminant SNPs was selected and used to develop a new run of the analysis. Again, the procedure was able to significantly identify the three breeds and to correctly assign new observations. These results suggest that an assay with the selected 48 SNP could be used to routinely track mono breed products.

Keywords: allocation method, bovine breeds, livestock products.

37 **Introduction**

38

39 The relevant concern of consumers about food quality has resulted in an increased importance of
40 products traceability in agriculture. Among the available identification methods, only DNA-based
41 techniques appear able to completely prevent from frauds. Microsatellite (Casellas *et al*. 2004; Orrù
42 *et al*. 2006; Dalvit *et al*. 2008) and AFLP markers (De Marchi *et al*. 2006; Negrini *et al*. 2007) have
43 been traditionally used for animal identification or parentage determination. More recently, a
44 different category of markers, the single nucleotide polymorphisms (SNP), have been proposed to
45 identify animals, breeds and their products. Compared to microsatellites, SNPs offer the advantage
46 they have lower rates of genotyping errors (Weller *et al.* 2006), are very abundant over the genome
47 (Heaton *et al.* 2005) and their analysis can be largely automatized.

48 At present, however, few studies have investigated the possible exploitation of SNPs for traceability
49 purposes. Orrù *et al.* (2009) tested 18 SNPs for their ability to identify individuals in six European
50 cattle breeds obtaining a probability to find two identical animals equal to 0.0765 out of one million
51 samples. Negrini *et al*. (2008) used a panel of 90 specifically selected SNPs to trace four European
52 protected indication beef products. Authors found a percentage of correct assignment ranging from
53 80% to 100%. Recently, Ramos *et al.* (2011) obtained 99% of correct assignment among five pig
54 breeds by using a SNP assay containing 193 breed specific markers.

55 All the above mentioned methods use a pool of pre-selected SNPs and suitable statistical techniques
56 to correctly assign individuals or animal derived foodstuffs. Essentially, two evaluation approaches
57 are used. The first is the deterministic and consists in finding SNPs with different allelic variants
58 fixed in the compared breeds (Paetkau *et al.* 1995). The second is the probabilistic and relies on
59 markers with typical allelic frequencies in different breeds. Statistical procedures as maximum
60 likelihood functions or Bayesian methods (Rannala & Mountain 1997) are therefore applied to

assign new observations to breeds. Several software packages are freely available to develop such analyses (Manel *et al.* 2005).

In this paper two multivariate statistical techniques were exploited to assess differences among three bovine breeds and to assign independent individuals to the proper group by using genomic data. The first objective was reached by using the canonical discriminant analysis (CDA) which extracts a set of linear combinations of the original variables able to maximize differences among predefined groups. The second was obtained by using the discriminant analysis (DA) which elaborates a discriminant function able to assign new observations to groups. Both techniques do not start from preselected variables, i.e. breed-specific SNPs. CDA and DA, analyze the correlation structure of SNPs in order to assess the difference among groups and assign new individuals. So, and this is one of the most important output of the CDA, a restricted pool of markers able to discriminate breeds is obtained at the end of the procedure.

Aims of the present work were a) to develop an efficient automated method for breed assignment and traceability purposes by using CDA and DA, b) to obtain a restricted pool of discriminant markers that could be used in traceability protocols.

Materials and methods

The data

Data consisted of 1,042 Holstein, 750 Brown Swiss and 480 Simmental bulls genotyped by using the Illumina 50K BeadChip (Matukumalli *et al*. 2009). Only markers located on the 29 autosomes were considered. SNP monomorphic, not in Hardy-Weinberg equilibrium, and with minor allele frequency lower than 5% were removed. This selective editing procedure obviously leads to discard SNPs fixed or typical for a specific breed, On the other hand, the aim of the present work is to use a

multivariate technique to detect a pool of highly discriminant markers based on their correlation structure and not, for example, on the occurrence of rare alleles. Finally, markers with more than 2.5% missing values were excluded. After data editing, the retained SNP were 38,450 for Holstein, 37,254 for Brown and 40,179 for Simmental, with 30,055 markers in common. The final matrix of data, however, still contained missing values. In this case, CDA and DA delete the corresponding rows, thus obtaining a very small data set. For this reason, missing data were imputed according to the most frequent genotype at each locus. Genotypes were finally coded as the number of copies of one SNP allele it carries, i.e. 0 (homozygous for allele A), 1 (heterozygous) or 2 (homozygous for allele B). Ten samples of 30 randomly selected bulls (10 for each breed) were generated and used as independent observations in the cross-validation procedure.

The Canonical discriminant analysis

The general objective of CDA is to distinguish among different populations by using a particular set of variables (Mardia *et al.* 2000). Unlike cluster analysis, in CDA the group to which each individual belongs is known. In this study CDA was applied to discriminate animals of three cattle breeds by using around 30K markers. Given the classification criterion (the breed), CDA derives a new set of variables, the canonical functions (CAN), which are linear combination of the original markers. The coefficients of the linear combination are the canonical coefficients (CC) which indicate the partial contribution of each original variable. When k-groups and m-variables are involved in the analysis, the maximum number of possible canonical functions is $p = \min(m; k\text{-}1)$. Being, in general, $m > k$, k-1 functions are derived. In the present work, being k-1=2, two canonical functions (CAN1 and CAN2) were derived.

106 The statistical significance in group separation can be expressed by means of the Mahalanobis'
107 distance and the corresponding Hotelling's T-square test (De Maesschalck *et al*., 2000). Groups are
108 declared significantly separated if the Hotelling's test shows a p-value less than 0.05. This test can
109 be developed only if the pooled (co)variance matrix of data is not singular. However, the visual
110 inspection of the CAN1 vs. CAN2 scatter-plot and the values of distances among groups can be
111 useful to asses if groups are separated. CDA and the related tests were developed by using the
112 CANDISC procedure implemented in the SAS-STAT software (SAS Institute Inc., Cary, NC,
113 USA). After differences among groups were assessed, the proc DISCRIM of SAS was used to
114 develop the DA. In this case, the canonical functions, applied to each animal, produced the
115 discriminant score: an individual is assigned to a particular group if its discriminant score is lower
116 than the cutoff-value obtained by calculating the weighted mean distance among group-centroids
117 (Mardia *et al.* 2000).

118

119 The CDA method for breed assignment

120 The matrix of data consisted of more than m = 30K SNP-variables and n = 2K animals. In this
121 condition, multivariate techniques became meaningless, being the rank of the extracted (co)variance
122 matrix ≤ n-1 (Dimauro *et al*. 2011). To overcome at least partially this problem, in genomic data
123 mining statistical analyses are often developed by chromosome (Macciotta *et al*. 2010). In the
124 present research, CDA was at first performed separately by each autosome. As a consequence, 29
125 CAN1 vs. CAN2 scatter-plots and 29 distance matrices were obtained. However, being the 29
126 pooled (co)variance matrices singular ($m>n$ in all chromosomes), the Mahalanobis' distance and the
127 related statistical test cannot be evaluated. Therefore, to obtain a pool of linearly independent
128 markers, canonical functions extracted for each chromosome were first ranked according to the CC
129 values. Then SNPs whose CC exceed an arbitrary fixed threshold were retained. So the final pool of

selected SNPs, besides linearly independent, were also the most discriminant. This markers were used to develop a genome wide CDA (GW-CDA) where both the Mahalanobis' distance and the Hotelling's test could be evaluated. Furthermore, the minimum subset of SNPs able to discriminate the three groups was also detected by using the same procedure applied to select the linearly independent SNPs.

To test the ability of the selected SNPs in assigning new animals to the proper breed, the DA was applied to the 10 cross-validation datasets previously generated. Moreover, the assignment test was also performed by using three independent algorithms included in the GeneClass2 software (Piry *et al*. 2004): the frequency-based method of Paetkau *et al*. (1995), the Bayesian-based methods of Rannala & Mountain (1997) and Baudouin & Lebrun (2000).

Results and discussion

CDA by chromosome

All CAN1 vs. CAN2 scatter plots displayed a clear separation among groups already at chromosome level, as shown in Figure1, where plots for BTAs 1 and 28 are reported. These chromosomes were chosen because they had the greater (BTA1) and the lower (BTA28) number of SNPs, respectively. Distances among breeds were different in the two chromosomes (figure 1). For example, the Euclidean distance between Holstein and the other two breeds on BAT28 was equal to 0.15 the corresponding distance on BTA1. The mean correlation value between distances among breeds and number of markers in each chromosome was around 0.75. This result clearly indicates that the multivariate description of a breed obtained by using genomic data produces, as expected, a greater separation among groups as the number available information (the markers) increases.

Distances between Brown and Simmental were lower than those for Holstein vs. Brown and Holstein vs. Simmental for all chromosomes. Similar results were obtained by Del Bo *et al.* (2001) who studied the genetic distances among 13 cattle breeds. Authors found a double distance among Holstein and the other two groups involved in the present study. A clear separation was also reported between Brown and Simmental.

Genome-wide CDA

In each chromosome, the threshold for the absolute value of CCs in CAN1 and CAN2 was arbitrarily fixed at 0.85 and 0.45 respectively. Different values were adopted for the two canonical functions because CC values in CAN1 were higher than in CAN2. A total of 1,836 SNPs were obtained and used to develop a GW-CDA. The resulting CAN1 vs. CAN2 scatter plot showed a clear separation of the three breeds (Figure 2) and, as in the by chromosome CDA, Holstein breed was markedly separated from the other two groups. The increase of distances between breeds for larger numbers of markers suggests that CDA is able to discriminate groups even if they are not markedly differentiated. It is worth remembering that the editing performed in this study has discarded rare alleles. Moreover, the selected SNPs used to develop the GW-DA gave 100% correct assignment of the new 30 observations in the 10 cross-validation datasets. This results clearly confirmed the goodness of the method in discriminating the three bovine breeds.

As at chromosome level, however, the **S** matrix of the 1,836 SNPs was singular. So, the number of markers was further reduced till to 340 linearly independent SNP-variables. The 340 SNP were then used to develop a new run of the GW-CDA. As in the previous cases, distances among breeds (table 1) showed a pattern like in CDA applied by chromosome. The Hotelling's test gave a highly

174 significant separation among breeds and GW-DA correctly assigned the animals in the cross-
175 validation datasets.

176 Finally, the selected 340 SNP-variables were reduced by deleting markers with lower CCs till to
177 reach the minimum number of markers able to highlight the existence of the groups. At the end, 48
178 of the most discriminant SNPs were retained and used in a new GW-CDA. A significant separation
179 among breeds was still obtained and the GW-DA was able to 100% assign animals in the 10 cross-
180 validation datasets. The same results were obtained with the GeneClass2 software, by using the
181 selected 48 SNPs. All animals were correctly assigned to the proper breed thus confirming the
182 ability of CDA in selecting markers able to discriminate the involved breeds.

183 As before, the CAN1 vs. CAN2 scatter plot (Figure 3) showed three well defined clusters with
184 Holstein clearly differentiated from the other two breeds. Markers and related CCs for each
185 canonical function are reported in table 2. Interesting considerations can be drawn by observing
186 Figure 3 and table 2. CAN1, which accounted for 92% of the total variability, shows very high CC
187 absolute values, ranging from 0,921 to 0,944. This result indicates that the associated markers
188 heavily affect the separation of Holstein from the other breeds. In figure 4a are displayed the
189 genotypic frequencies for SNP having the negative CC. It can be clearly noticed that the
190 predominant homozygous genotype in Holstein is the opposite of the other breeds. For example, BB
191 is the most frequent genotype in Holstein whereas in Simmental and Brown is the most rare. A
192 reversed pattern is shown for SNPs having positive CCs (figure 4b). For CAN2, which accounted
193 only for the 8% of the total variability, the differences among the genotypic frequencies are less
194 marked and, therefore were not reported.

195

196 **Conclusions**

The study demonstrated that the canonical discriminant analysis was able to efficiently distinguish the three breeds involved in the research by using genomic data, also at chromosome level. The high correlation (0.75) between the number of SNPs in a chromosome and the distance among breeds suggested that the more markers are involved the more efficiently groups are discriminated. The subsequent GW-CDA developed by using a reduced number of markers (1,836), chosen among most discriminants, confirmed the ability of the method in separating groups. These results suggested that if really different breeds are under study, even if not highly differentiated, a clear separation could be reached by enlarging the number of SNPs involved in the analysis. however, further analyses involving other breeds should be carried out to confirm this hypothesis. The Hotelling's statistical test evaluated in the GW-CDA developed by using 340 linearly independent SNPs indicated an highly significant difference among breeds, thus confirming the hypothesis that the three cattle populations can be differentiated by using genomic variables. The technique does not require a pool of preselected markers being the detection of the most discriminant markers one of the expected outputs. However, to assess the difference among breeds by using the Hotelling's test, around 2,000 genotyped animals are required. Finally, 48 SNPs were able to separate groups and, by using the DA, new observations were 100% correctly assigned. Moreover, the assignment tests developed by using an independent software as GeneClass2, confirmed the ability of CDA in selecting pool of discriminant markers. The selected 48 markers could be used to create an assay that could be routinely applied to trace milk, meat or other animal products derived from the three breeds involved in the study.

Acknowledgements Work funded by the Italian Ministry of Agriculture (grant SELMOL and Innovagen)

222 **References**

223 Baudouin L. & Lebrun P. (2000) An operational Bayesian approach for the identification of
224 sexually reproduced cross-fertilized populations using molecular markers. *Acta*
225 *Horticulturae* 546, 81–93.

226 Casellas J., Jimenez N., Fina M., Tarres J., Sanchez A. & Piedrafita J. (2004) Genetic diversity
227 measures of the bovine Alberes breed using microsatellites, variability among herds and
228 types of coat colour. *Journal of Animal Breeding and Genetics* 121, 101–10.

229 Dalvit C., De Marchi M., Targhetta C., Gervaso M. & Cassandro M. (2008) Genetic traceability of
230 meat using microsatellite markers. *Food Research International* 41, 301–7.

231 Del Bo L., Polli M., Longeri M., Ceriotti G., Looft C., Barre-Dire A., Golf G. & Zanotti M. (2001)
232 Genetic diversity among some cattloe breeds in the Alpine area. *Journal of Animal Breeding*
233 *and Genetics* 118, 317-25

234 De Maesschalck R., Jouan-Rimbaud D. & Massart D. L. (2000) The Mahalanobis distance.
235 *Chemometrics and Intelligent Laboratory Systems* 50, 1–18

236 De Marchi M., Dalvit C., Targhetta C. & Cassandro M. (2006) Assesing genetic diversity in
237 indigenous Veneto chicken breeds using AFLP markers. *Animal Genetics* 37, 101–105.

238 Dimauro C., Cellesi M., Pintus M. A. & Macciotta N. P. P. (2011) The impact of the rank of marker
239 variance–covariance matrix in principal component evaluation for genomic selection
240 applications. *Journal of Animal Breeding and Genetics* 128, 440–5

241 Heaton M. P., Keen J. E., Clawson M. L., Harhay G. P., Bauer N., Shultz C., et al. (2005) Use of
242 bovine single nucleotide polymorphism markers to verify sample tracking in beef
243 processing. *Journal of the American Veterinary Medical Association* 226, 1311–4.

244 Macciotta N. P. P. , Gaspa G., Steri R. , Nicolazzi E. L., Dimauro C., Pieramati C. & Cappio-
245 Borlino A. (2010) Using eigenvalues as variance priors in the prediction of genomic
246 breeding values by principal component analysis. *Journal of Dairy Science* 93, 2765–74

247 Manel S., Gaggiotti O. E. & Waples R. S. (2005) Assignment methods: Matching biological
248 questions with appropriate techniques. *Trends in Ecology & Evolution* 20, 136–42.

249 Mardia K. V, Kent J. T. & Bibby J. M. (2000) *Multivariate analysis*. Academic Press, London

250 Matukumalli L.K., Lawley C.T., Schnabel R.D. et al. (2009) Development and characterization of a
251 high density SNP genotyping assay for cattle. *PLoS ONE* 4, e5350.

252 Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P., Lenstra J.A. & Ajmone-
253 Marsan P. (2007) Breed assignment of Italian cattle using biallelic AFLP markers. *Animal*
254 *Genetics* 38, 147–53.

255 Negrini R., Nicoloso L., Crepaldi P., Milanesi E., Colli L., Chegdani F., Pariset L., Dunner S.,
256 Leveziel H., Williams J. L. & Ajmone Marsan P. (2008) Assessing SNP markers for
257 assigning individuals to cattle populations. *Animal Genetics* 40, 18–26

258 Orrù L., Napolitano F., Catillo G. & Moioli B. (2006) Meat molecular traceability: How to choose
259 the best set of microsatellites? *Meat Science* 72, 312–7

260 Orrù L., Catillo G., Napolitano F., De Matteis G., Scatà M.C., Signorelli F. & Moioli B. (2009)
261 Characterization of a SNPs panel for meat traceability in six cattle breeds. *Food Control* 20,
262 856–60

263 Paetkau D., Calvert W., Stirling I. & Strobeck C. (1995). Microsatellite analysis of population
264 structure in Canadian polar bears. *Molecular Ecology* 4, 347–54.

265 Piry S., Alapetite A., Cornuet J.M., Paetkau D., Baudouin L. & Estoup A. (2004) GeneClass2: a
266 software for genetic assignment and first-generation migrant detection. *Journal of Heredity*
267 95, 536–9.

268 Ramos A. M., Megens H. J., Crooijmans R. P. M., Schook L. B. & Groenen M. A. M. (2011)
269 Identification of high utility SNPs for population assignment and traceability purposes in the
270 pig using high-throughput sequencing. *Animal Genetics* 42, 613–20

271 Rannala B. & Mountain J. (1997) Detecting immigration by using multilocus genotypes.
272 *Proceedings of the National Academy of Sciences* 94, 9197–201.

273 Weller J. I., Seroussi E. & Ron M. (2006) Estimation of the number of genetic markers required for
274 individual animal identification accounting for genotyping errors. *Animal Genetics* 37, 387–
275 9.

276

277

278 **Table 1** Mahalanobis' distances among group centroids of breeds and, in bracket, the Hotelling's
279 test of significance evaluated by using 340 linearly independent SNPs

	Brown	Simmental
Simmental	301 (<0.0001)	
Holstein	4300 (<0.0001)	3574 (<0.0001)

280

281

282

283

284

285 **Table 2** Canonical coefficients (CC), in the two canonical functions (CAN1 and CAN2), for the
286 most 48 discriminant markers selected among SNPs belonging to the Illumina BovineSNP50 v2
287 BeadChip

SNP name	BTA	CC (CAN1)	SNP name	BTA	CC (CAN2)
BTB-01524285	5	0.944	Hapmap56688-rs29025335	6	-0.671
ARS-BFGL-NGS-116089	15	0.941	ARS-BFGL-NGS-100916	6	-0.666
Hapmap51971-BTA-18711	11	0.936	ARS-BFGL-NGS-103634	18	-0.664
BTB-01648149	3	0.936	Hapmap30962-BTC-032558	6	-0.651
BTA-23857-no-rs	12	0.933	ARS-BFGL-NGS-41271	20	-0.648
BTB-01267305	5	0.932	ARS-BFGL-NGS-108820	6	-0.645
BTA-73563-no-rs	5	0.931	BTB-00049653	1	-0.640
BTA-79188-no-rs	1	0.930	Hapmap27224-BTA-161106	6	-0.640
ARS-BFGL-NGS-3048	29	0.929	ARS-BFGL-NGS-67658	6	-0.640
BTB-00498059	12	0.928	BTB-00259302	6	-0.639
Hapmap33485-BTA-144281	6	0.928	Hapmap54879-rs29017018	6	-0.635
Hapmap55512-rs29011234	26	0.928	Hapmap52160-rs29020798	6	-0.627
ARS-BFGL-NGS-22403	16	-0.921	ARS-BFGL-NGS-20141	7	0.633
BTA-58999-no-rs	24	-0.922	BTA-37834-no-rs	5	0.636
UA-IFASA-3757	13	-0.922	BTA-110240-no-rs	6	0.636
BTB-00506196	12	-0.922	Hapmap42715-BTA-87995	6	0.643
BTB-00951350	27	-0.925	Hapmap57799-rs29012894	11	0.643
BTB-00506214	12	-0.926	ARS-BFGL-BAC-33135	18	0.650
ARS-BFGL-NGS-36907	26	-0.928	Hapmap50117-BTA-81807	6	0.650
BTB-00146014	3	-0.928	Hapmap44452-BTA-22099	6	0.681
Hapmap44270-BTA-67318	9	-0.928	Hapmap33128-BTC-041916	6	0.766
BTB-00178642	4	-0.928	Hapmap26269-BTC-041695	6	0.782
BTA-18115-no-rs	2	-0.937	ARS-BFGL-NGS-38827	6	0.785
Hapmap51008-BTA-62521	27	-0.943	Hapmap27692-BTC-042876	6	0.787

288

289

290

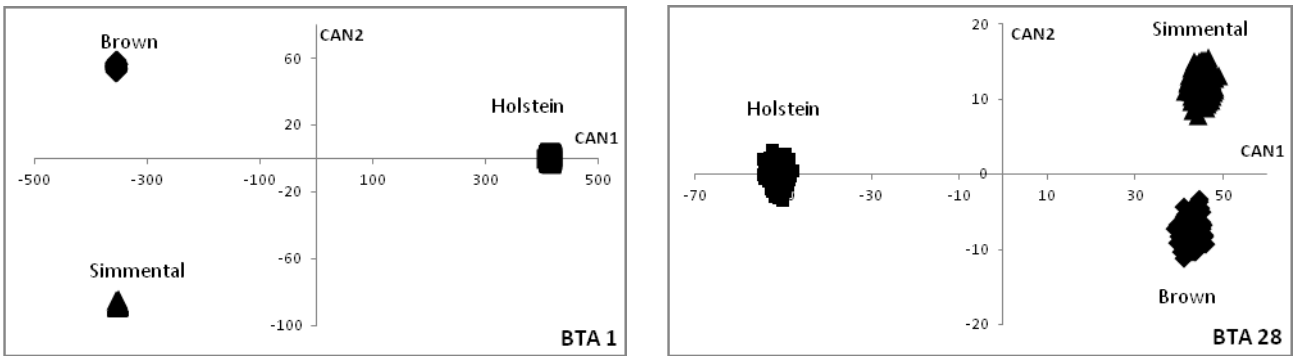
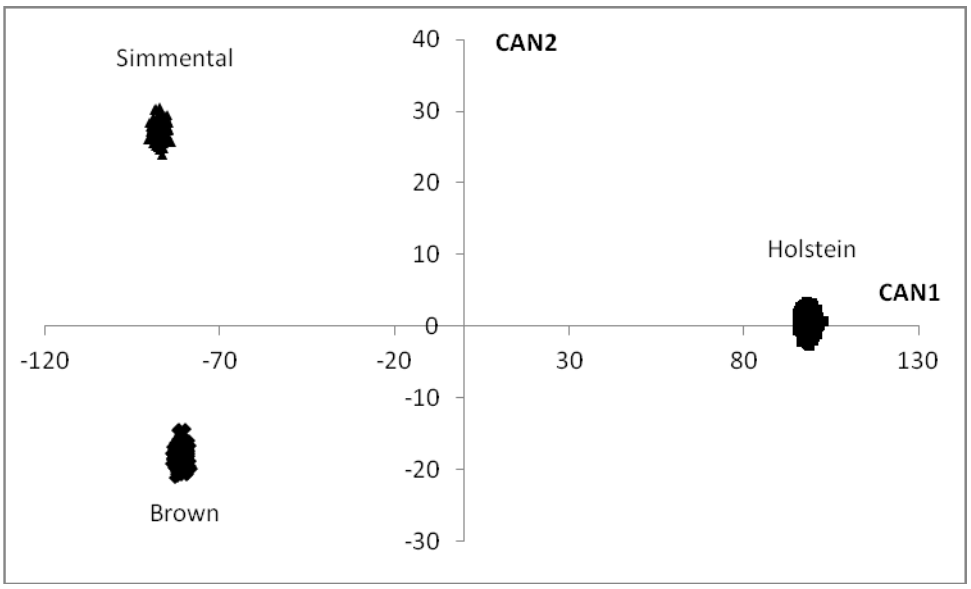


Figure 1 Graph of the two canonical functions (CAN1 and CAN2) obtained in a canonical discriminant analysis applied to BTA1 and BTA28, the two chromosomes with the greater and the lower number of SNP-variables, respectively.

298

299



300

301 **Figure 2** Graph of the two canonical functions (CAN1 and CAN2) obtained in a genome wide

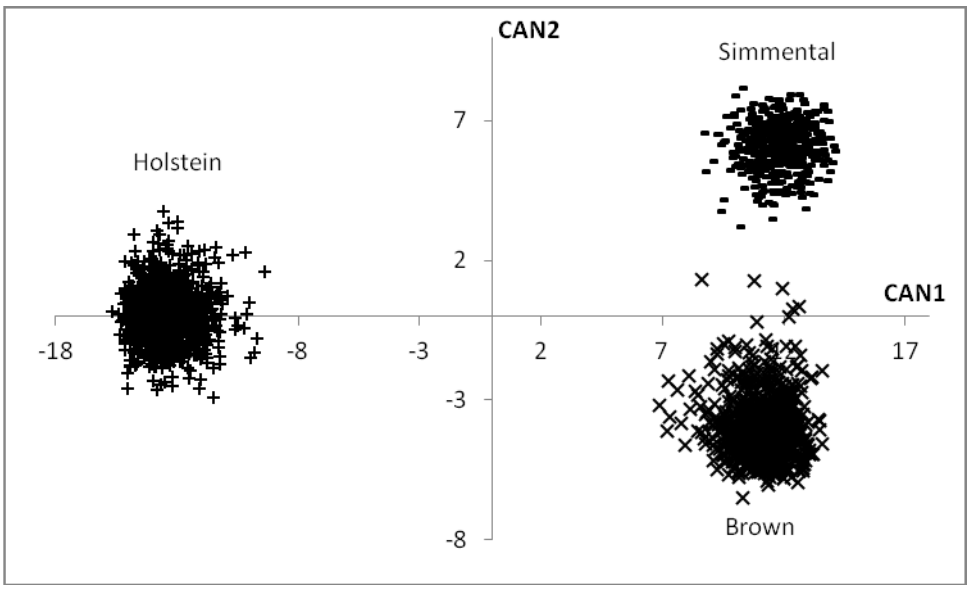
302 canonical discriminant analysis by using a restricted number (1836) of SNP-variables

303

304

305

306



307

308 **Figure 3** Graph of the two canonical functions (CAN1 and CAN2) obtained in a genome wide
309 canonical discriminant analysis by using a restricted number (48) of linearly independent SNP-
310 variables.

311

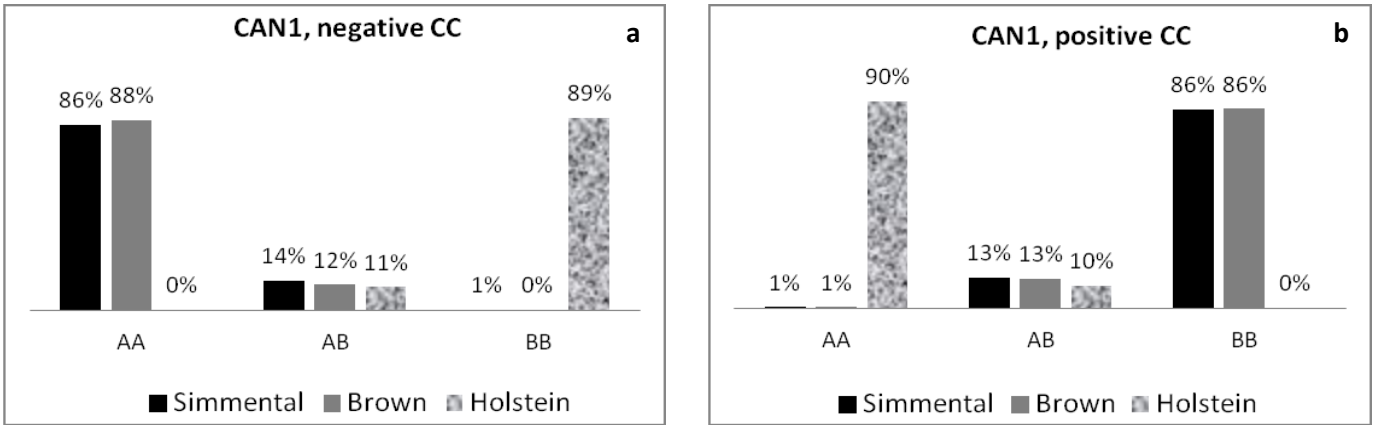
312

313

314

315

316



317 **Figure 4** Genotypic frequencies for 48 highly discriminant SNPs for negative (a) and positive (b)

318 canonical coefficients (CC) in the first canonical function (CAN1)

319